Interpreting Item Analysis Data

Provided by:

Test Validation & Construction Unit

California State Personnel Board

Interpreting Item Analysis Data

This document describes and explains the information that is found on the State Personnel Board's item analysis printout. A more detailed explanation of item analysis data can be found in *Section 5360* of the State Personnel Board's Selection Manual.

The item analysis printout is a statistical summary of the results of a written exam administered to a group of examinees. As such, this printout is used to evaluate the effectiveness of the exam. More specifically, the information included on the item analysis printout is used to identify miskeyed or malfunctioning items, as well as items that simply do not discriminate between the better candidates and the poorer candidates. Item analysis information is also useful when revising items for future administrations of the exam.

It is important that the item analysis for an exam be run and interpreted by exam segment. Running the item analysis by exam segment is necessary because the statistics used to estimate exam reliability are based on the assumption that exam content is homogeneous. An entire written exam is typically composed of several heterogeneous segments. Therefore, as an aid in the proper interpretation of the item analysis data, the item analysis report will automatically compute all item analysis statistics by exam segment.

The item analysis printout contains statistical information for each individual item within an exam segment, as well as summary information for the exam segment as a whole. The remainder of this document will illustrate and explain these statistics.

Item Statistics

As depicted in the following table, the item statistics information on the item analysis report is divided into four columns: Item Number, Frequency, Percent, and Item Discrimination.

Item Number	Frequency						Percent						Item Discrimination					
	A	B	C	D	E	O	A	B	C	D	E	O	A	B	C	D	E	O
21	*						*						*					
UPPER	134			1			99			1								
MIDDL	205	2	3	21			89	1	1	9								
LOWER	79	11	14	31			59	8	10	23								
TOTAL	418	13	17	53			83	3	3	11			.42	.19-	.21-	.29-		

Item Number

Item number is the exam question being analyzed. In the example below, 21, found directly below the heading *Item Number*, corresponds to question 21 in the exam. All of the information in the table pertains to question 21 in the exam.

Item Number	Frequency						Percent						Item Discrimination					
	A	B	C	D	E	O	A	B	C	D	E	O	A	B	C	D	E	O
21	*						*						*					
UPPER	134			1			99			1								
MIDDL	205	2	3	21			89	1	1	9								
LOWER	79	11	14	31			59	8	10	23								
TOTAL	418	13	17	53			83	3	3	11			.42	.19-	.21-	.29-		

Upper, Middl, Lower

The item analysis printout splits the candidates into three groups based upon their scores on the exam segment. In other words, if candidate scores on the exam segment were arrayed in descending order, the 27% of the candidates with the highest scores on the segment would be placed in the UPPER group, the 27% percent of the candidates with the lowest scores on the segment would be placed in the LOWER group, and the remaining 46% of the candidates would be placed in the middle group. The middle group is identified as MIDDL on the item analysis printout. TOTAL is simply the total group of candidates regardless of their score on the exam segment.

UPPER = top 27% of candidates on the exam segment
MIDDL = middle 46% of candidates on the exam segment
LOWER = bottom 27% of candidates on the exam segment

Frequency

The item analysis printout provides a frequency count of the candidates within the Upper, Middle, and Lower groups who selected answer choice A, B, C, D, or E. Candidates who failed to answer the question are identified in the column labeled O (for omit).

For question 21 in the above example, 134 candidates in the Upper group selected answer choice A, and one candidate in the Upper group selected D. The frequency counts are also provided for the Middle and Lower groups. Note that a Total frequency count is also provided for each response option. This is simply the sum of each column.

The keyed response, or correct answer, is identified by an asterisk (*). In the example below, A is the keyed response.

Item Number	Frequency						Percent						Item Discrimination					
	A	B	C	D	E	O	A	B	C	D	E	O	A	B	C	D	E	O
21	*						*						*					
UPPER	134			1			99			1								
MIDDL	205	2	3	21			89	1	1	9								
LOWER	79	11	14	31			59	8	10	23								
TOTAL	418	13	17	53			83	3	3	11			.42	.19-	.21-	.29-		

Percent

Percent identifies the percentage of candidates within the Upper, Middle, and Lower groups who selected answer choice A, B, C, D, E, or who failed to answer the question.

For question 21 in the above example, 99% of the candidates in the Upper group selected answer choice A, the correct answer, and 1% of the candidates in the Upper group selected D. Percent is provided for the Middle and Lower groups in a similar manner. Note that a Total percent value is also provided. Total percent is the percentage of all candidates who selected each response option. For question 21, 83% of the candidates selected A as the correct answer.

The percentage of candidates selecting the correct answer is used as an index of the item difficulty. For question 21, .83 is the item difficulty. An item difficulty index between .40 and .60 is desired. That is, ideal items are items that approximately 40 to 60 percent of the candidates answer correctly.

Item Discrimination

There are several item discrimination indices which provide a measure of how well an exam item discriminates between the better candidates and the poorer candidates. The item discrimination statistic used by the State Personnel Board's on-line system is the point-biserial correlation.

The point-biserial correlation indicates the direction and strength of the relationship between a pair of scores. In the case of an item analysis, the point-biserial correlation is a correlation between segment score (a continuous variable) and item score (a dichotomous variable).

- **Direction**

  A positive coefficient for a response option indicates that candidates who performed well on the exam segment also selected that particular response option. The point-biserial correlation coefficient for the key, or correct answer, should always be positive. When this is the case, the coefficient indicates that those candidates who performed well on the exam segment selected the keyed response.

  A negative coefficient for a response option indicates that candidates who performed well on the exam segment did NOT choose that particular response option. The point-biserial correlation coefficients for the distractors, or incorrect response options, should always be negative. A negative point-biserial correlation coefficient for the keyed response is an indication that the item is problematic. The problem may simply be that the item has been miskeyed, or the item may be ambiguous, confusing, or malfunctioning for some other reason. If the item has been miskeyed, the key should be corrected prior to finalizing candidate scores. If the item has been keyed correctly but is malfunctioning, the item should probably be deleted prior to finalizing scores.

- **Strength**

  Strength refers to the size or magnitude of the point-biserial correlation coefficient. A point-biserial correlation coefficient can range in value between –1.00 and +1.00. A high point-biserial coefficient for the keyed response is desired since this indicates that those candidates who did well on the exam segment are getting the item correct, while those candidates who did poorly on the exam segment are getting the item wrong. Consequently, the item is doing a good job of discriminating between the better and the poorer candidates.

  The greater the number of candidates in the Upper group who correctly answer the item, the higher the point-biserial coefficient will be. The better items will be those which are answered correctly by all of the candidates in the Upper group, and none of the candidates in the lower group. Point-biserial correlation coefficients for the keyed response of .40 and above are considered to be very good. Coefficients in the range of .30 to .39 are good. A point-biserial correlation coefficient of less than .30 for the keyed response indicates that the item is not doing an optimal job of discriminating between the better and poorer candidates.

  Point-biserial correlation coefficients for the distractors of -.30 and below are considered to be very good. Coefficients of -.20 to -.29 are

good, while coefficients greater than -.20 may be indicative of a problematic distractor.

Segment Statistics

The item analysis printout also includes statistics for each exam segment. Segment statistics can be found on the item analysis printout directly following the last set of item statistics in the segment. An example illustrating the segment statistics is presented below.

Segment R .89	Mean 37.99	S D 6.57	Mean Item Difficulty 55	No. of Items 46

Segment R

Segment R refers to the reliability of the exam segment. Reliability is defined as the extent to which scores achieved on that exam segment are precise or stable indicators of the candidates' true level of knowledge or skill. When reliability is low, there is an increased chance of accepting candidates who cannot perform the job, or eliminating candidates who can perform the job.

Exam segment reliability is calculated using Cronbach's Coefficient Alpha. Coefficient Alpha can range in value between 0.0 and +1.0. The greater the coefficient, the more reliable the exam segment. Coefficients greater than .80 are desirable.

Reliability statistics are based on the assumption that exam content is homogeneous. An entire written exam is usually composed of several heterogeneous segments. Therefore, only the reliability statistics for the individual exam segments are reported in the item analysis printout

Mean

The mean indicates the average candidate score in the distribution of scores. The mean is a measure of central tendency. Measures of central tendency are indicators of a distribution's average or typical score.

S D -- Standard Deviation

The standard deviation, identified as S D on the item analysis printout, is a measure of the dispersion of the exam scores about the mean score. In effect, the standard deviation is the average of the difference of the candidates' scores from the mean score. The larger the standard deviation, the more the scores differ from each other. The standard deviation will be 0.00 if all of the scores are the same. A relatively large

standard deviation is desired for an exam segment. As a rule of thumb, a good standard deviation (that is, one that provides for a good distribution of scores) is 10% of the length of the exam segment or greater.

Mean Item Difficulty

Mean item difficulty is the average item difficulty for the exam segment. A mean item difficulty between .40 and .60 is desirable. Mean item difficulty values of this magnitude indicate that, on average, 40 to 60 percent of the candidates are getting a given item correct.

No. of Items

Number of items indicates the number of items within that exam segment. In order to achieve acceptable exam segment reliabilities, each segment should consist of approximately 30 or more items.